

The support reduction algorithm for computing nonparametric function estimates in mixture models

Piet Groeneboom, Geurt Jongbloed*, Jon A. Wellner

February 1, 2008

Abstract

Vertex direction algorithms have been around for a few decades in the experimental design and mixture models literature. We briefly review this type of algorithms and describe a new member of the family: the support reduction algorithm. The support reduction algorithm is applied to the problem of computing nonparametric estimates in two inverse problems: convex density estimation and the Gaussian deconvolution problem. Usually, VD algorithms solve a finite dimensional (version of the) optimization problem of interest. We introduce a method to solve the true infinite dimensional optimization problem.

1 Introduction

During the past decades emphasis in statistics has shifted from the study of parametric models to that of semi- or nonparametric models. A big advantage of these latter models is their flexibility and ability to ‘let the data speak for itself’. However, also problems that were not usually crucial in the parametric case, turn out to be difficult in the semiparametric situation. The asymptotic distribution theory of estimators is one of these problems. The multivariate central limit theorem and the delta method give the answer to many questions regarding asymptotic distribution theory in the parametric setting. For the semiparametric situation, such ‘basic tools’ are not available. Another problem that is usually easier to solve in parametric models is the problem of computing M -estimators that are defined as minimizer of a random criterion function. In a parametric model often estimates can be computed explicitly or computed using some numerical technique for solving (low dimensional) convex unconstrained optimization problems like steepest descent or Newton. In semiparametric models, the computational issues often boil down to high dimensional constrained optimization problems.

Apart from algorithms that are known from the general theory of optimization, algorithms have been designed within the field of statistics that are particularly useful

*Research supported by a grant from the Haak Bastiaanse Kuneman foundation of the Vrije Universiteit.

in certain statistical applications. Perhaps the best known example of this type is the *Expectation Maximization* (EM) algorithm of DEMPSTER, LAIRD AND RUBIN (1977), that is designed to compute maximum likelihood estimates based on incomplete data. Another example is the *iterative convex minorant* algorithm that is introduced in GROENEBOOM AND WELLNER (1992) and further studied in JONGBLOED (1998). That algorithm is based on techniques known from the theory of isotonic regression as can be found in ROBERTSON, WRIGHT, AND DYKSTRA (1988) and can be used to compute nonparametric estimators of distribution functions in semiparametric models. Another class of algorithms that falls within this framework is the class of *vertex direction* (VD) algorithms.

In section 2 we introduce the general structure of VD algorithms and mixture models where VD algorithms can be used to compute nonparametric function estimates. Two specific examples of these mixture models will be considered in subsequent sections: estimating a convex decreasing density and estimating a mixture of unit variance normal distributions.

In section 3, we introduce the support reduction algorithm as a specific member of the VD family of algorithms. This algorithm essentially replaces the original infinite dimensional constrained optimization problem by a sequence of finite dimensional unconstrained optimization problems. The algorithm is designed to keep the dimension of these subproblems as low as possible. For a specific type of statistical models, the algorithm seems to be a good candidate to compute sensible estimators. These are problems that are difficult from the asymptotic statistical point of view in the sense that the convergence rate of the estimator is relatively low.

All VD algorithms have to deal with a problem of minimizing a “directional derivative” function over some set of parameters. There are some variants of these functions. For quadratic objective functions, we will describe an alternative directional derivative function in section 4 that takes more local information of the objective function into account.

The directional derivative function (and our alternative) are usually nonconvex functions on a continuum of parameters. Usually the associated nonconvex minimization problem is circumvented by considering a fine grid within the parameter space and minimizing the function only over that grid. In section 5 we propose a method of “leaving the grid”, tackling the infinite dimensional optimization problem rather than the finite dimensional approximation.

Section 6 is devoted to least squares estimation within a mixture model. The general procedure is given and for the problem of estimating a convex and decreasing density based on a sample from it, will be considered in detail. In that situation the support reduction algorithm boils down to what is called the iterative cubic spline algorithm in GROENEBOOM, JONGBLOED AND WELLNER (2001B).

In section 7 the general problem of computing a maximum likelihood estimate within a mixture model will be addressed. A Newton procedure based on the support reduction algorithm will be described. The normal deconvolution problem will serve as example to illustrate the general approach.

2 Vertex direction-type algorithms

Consider the following type of optimization problem

$$\text{minimize } \phi(f) \quad \text{for } f \in C \quad (2.1)$$

where ϕ is a convex function defined on (a superset of) a convex set of functions C . We assume throughout that ϕ has a unique minimizer over C .

Assumption A1: ϕ is a convex function on C such that for each $f, g \in C$ where ϕ is finite, the function $t \mapsto \phi(f + t(g - f))$ is continuously differentiable for $t \in (0, 1)$.

Now define, for each $f \in C$ and h a function such that for some $\epsilon > 0$, $f + \epsilon h \in C$,

$$D_\phi(h; f) = \lim_{\epsilon \downarrow 0} \epsilon^{-1} (\phi(f + \epsilon h) - \phi(f))$$

Note that this quantity exists (possibly equal to ∞) by convexity of ϕ . As we will see, a choice often made for h is $h = g - f$ for some arbitrary $g \in C$. In that case we have

$$D_\phi(g - f; f) = \lim_{\epsilon \downarrow 0} \epsilon^{-1} (\phi(f + \epsilon(g - f)) - \phi(f))$$

The following simple but important result gives necessary and sufficient conditions for \hat{f} to be the solution of (2.1).

Lemma 2.1 *Suppose that ϕ satisfies A1. Then*

$$\hat{f} = \operatorname{argmin}_{f \in C} \phi(f) \quad \text{if and only if} \quad D_\phi(g - \hat{f}; \hat{f}) \geq 0 \quad \text{for all } g \in C.$$

Proof: First we prove \Rightarrow . Suppose $\hat{f} = \operatorname{argmin}_{f \in C} \phi(f)$ and choose $g \in C$ arbitrarily. Then, for $\epsilon \downarrow 0$

$$0 \leq \epsilon^{-1} (\phi(\hat{f} + \epsilon(g - \hat{f})) - \phi(\hat{f})) \downarrow D_\phi(g - \hat{f}; \hat{f})$$

Now \Leftarrow . For arbitrary $g \in C$, write τ for the convex function $\epsilon \mapsto \phi(\hat{f} + \epsilon(g - \hat{f}))$ and note that

$$\phi(g) - \phi(\hat{f}) = \tau(1) - \tau(0) \geq \tau'(0+) = D_\phi(g - \hat{f}; \hat{f}) \geq 0.$$

□

Consider now the case where C is the convex hull of a class of functions

$$\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\}, \quad (2.2)$$

in the sense that

$$C = \operatorname{conv}(\mathcal{F}) = \left\{ g = \int_{\Theta} f_\theta d\mu(\theta) : \mu \text{ probability measure on } \Theta \right\}. \quad (2.3)$$

Here are two examples of *mixture models* that fall within this framework. These examples will reappear in subsequent sections.

Example 1. (*convex decreasing density*)

The class of convex decreasing densities on $[0, \infty)$ has representation (2.3) with

$$f_\theta(x) = \frac{2(\theta - x)}{\theta^2} 1_{(0, \theta)}(x), \quad \theta > 0.$$

It is obvious that any (positive) mixture of these functions is convex and decreasing. Since the mixing measure is a probability measure, it also follows that the mixture is a probability density. To see that any convex and decreasing density can be written as mixture of f_θ 's, note that the measure defined by $d\mu(\theta) = \frac{1}{2}\theta^2 df'(\theta)$ gives

$$\int_0^\infty f_\theta(x) d\mu(\theta) = \int_x^\infty \frac{2(\theta - x)}{\theta^2} d\mu(\theta) = \int_x^\infty (\theta - x) df'(\theta) = f(x).$$

Situations where the problem of estimating a convex and decreasing density based on a sample from it is encountered, can e.g. be found in HAMPEL (1987) and LAVEE, SAFRIE, AND MEILIJSON (1991). \square

Example 2. (*mixture of unit variance normals*)

The *Gaussian deconvolution problem* as considered in e.g. GROENEBOOM AND WELLNER (1992), entails estimation of a density (and associated mixing distribution) that belongs to the convex hull of the class of normal densities with unit variance:

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}.$$

\square

In the examples just considered, usually one has a sample X_1, X_2, \dots, X_n from the unknown density $f \in C$, and wants to estimate the underlying density f based on that sample. In this paper we consider two types of nonparametric estimators: least squares (LS) estimators and maximum likelihood (ML) estimators.

Least Squares estimation.

We define a least squares estimate of the density in C as minimizer of the function

$$\phi(f) = \frac{1}{2} \int_0^\infty f(t)^2 dt - \int_0^\infty f(t) d\mathbb{F}_n(t) \quad (2.4)$$

over the class C . Here \mathbb{F}_n is the empirical distribution function of the sample.

The reason for calling this estimator a LS estimator, is the following heuristic. For any (arbitrary) square integrable density estimate \tilde{f}_n of f_0 , one can define the LS estimate as minimizer of the function

$$f \mapsto \frac{1}{2} \int_0^\infty (f(t) - \tilde{f}_n(t))^2 dt = \frac{1}{2} \int_0^\infty f(t)^2 dt - \int_0^\infty f(t) \tilde{f}_n(t) dt + c_{\tilde{f}_n} \quad (2.5)$$

over the class C . It is seen that, as far as minimization over f is concerned, (2.5) only depends on the density \tilde{f}_n via its distribution function. The objective function in (2.4) is

obtained by taking the empirical distribution function for this estimator, so we take formally $\tilde{f}_n(t) dt = d\mathbb{F}_n(t)$ in (2.5). Note that for objective function (2.5)

$$D_\phi(h; f) = \lim_{\epsilon \downarrow 0} \epsilon^{-1} (\phi(f + \epsilon h) - \phi(f)) = \int h(x) f(x) dx - \int h(x) d\mathbb{F}_n(x).$$

Maximum Likelihood estimation.

As maximum likelihood estimate we define the minimizer of the function

$$\phi(f) = - \int \log f(x) d\mathbb{F}_n(x)$$

over the class of densities C . Note that for this function

$$D_\phi(h; f) = \lim_{\epsilon \downarrow 0} \epsilon^{-1} (\phi(f + \epsilon h) - \phi(f)) = - \int \frac{h(x)}{f(x)} d\mathbb{F}_n(x).$$

For both objective functions ϕ , the function D_ϕ has the linearity property stated below.

Assumption A2: the function ϕ has the property that for each $f \in C$ and $g = \int_\Theta f_\theta d\mu_g(\theta) \in C$,

$$D_\phi(g - f; f) = \int_\Theta D_\phi(f_\theta - f; f) d\mu_g(\theta). \quad (2.6)$$

Under this additional assumption, the nonnegativity condition in lemma 2.1 that has to hold for each $g \in C$, may be restricted to functions $g \in \mathcal{F}$.

Lemma 2.2 *Suppose that $C = \text{conv}(\mathcal{F})$ with \mathcal{F} as in (2.2) and that ϕ satisfies **A1** and **A2**. Then*

$$\hat{f} = \operatorname{argmin}_{f \in C} \phi(f) \quad \text{if and only if} \quad D_\phi(f_\theta - \hat{f}; \hat{f}) \geq 0 \quad \text{for all } \theta \in \Theta.$$

Proof: Follows immediately from lemma 2.1, the fact that $f_\theta \in C$ and (2.6) □

For the situation of Lemma 2.2, there is a variety of algorithms to solve (2.1) that can be called ‘of vertex direction (VD) type’. A common feature of VD algorithms is that they consist of two basic steps. Given a current iterate f , find a value of θ such that $D_\phi(f_\theta - f; f)$ is negative. (If such a value cannot be found, the current iterate is optimal!) This means that travelling from the current iterate in the direction of f_θ would (initially) decrease the value of the function ϕ .

Having found such a feasible profitable direction from the current iterate, the next step is to solve some low-dimensional optimization problem to get to the next iterate.

The original algorithm, proposed by FEDOROV (1972) and WYNN (1970) in the context of computing an optimal design, as well as the algorithm proposed by SIMAR (1976) (for computing the maximum likelihood estimate of the mixing distribution in a Poisson mixture) that we will come back to later, implement the first step as follows. Given the current f , find $\hat{\theta}$ corresponding to the minimizer of $D_\phi(f_\theta - f; f)$ over Θ .

FEDOROV (1972) and WYNN (1970) then propose to take as new iterate the function

$$g = (1 - \hat{\epsilon})f + \hat{\epsilon}f_{\hat{\theta}}$$

where $\hat{\epsilon}$ is given by

$$\hat{\epsilon} = \operatorname{argmin}_{\epsilon \in [0,1]} \phi((1 - \epsilon)f + \epsilon f_{\hat{\theta}}).$$

In words, the next iterate is the optimal convex combination of the current iterate and the most promising vertex in terms of the directional derivative. It is clear that usually the next iterate has one more support point than the current iterate.

The *vertex exchange algorithm* as proposed in BÖHNING (1986), not only uses the parameter $\hat{\theta}$ corresponding to the minimizer of $D_\phi(f_\theta - f; f)$, but also the maximizer $\check{\theta}$ of $D_\phi(f_\theta - f; f)$ restricted to the support points of the current iterate to get the direction. Denote by $\mu_f(\{\check{\theta}\})$ the mass assigned by the mixing distribution corresponding to f to $\check{\theta}$. Then the direction given by the algorithm is $f + \mu_f(\{\check{\theta}\})(f_{\hat{\theta}} - f_{\check{\theta}})$. The new iterate becomes

$$f + \hat{\epsilon} \mu_f(\{\check{\theta}\})(f_{\hat{\theta}} - f_{\check{\theta}})$$

where

$$\hat{\epsilon} = \operatorname{argmin}_{\epsilon \in [0,1]} \phi(f + \epsilon \mu_f(\{\check{\theta}\})(f_{\hat{\theta}} - f_{\check{\theta}})).$$

If $\hat{\epsilon} = 1$, the point $\check{\theta}$ is eliminated from the support of the current iterate, and the mass assigned to $\check{\theta}$ by the ‘old’ mixing distribution, is moved to the new point $\hat{\theta}$. It is clear that in this algorithm the number of support points of the iterate can increase by one, remain the same, but also decrease by one during one iteration (if $\hat{\epsilon} = 1$ and $\check{\theta}$ already belongs to the support). In specific examples, the number of support points of the solution \hat{f} is known to be smaller than a constant N which only depends on the data (and is known in advance). In the context of random coefficient regression models, MALLET (1986) proposes to restrict all iterates to having at most N support points.

Another variation on the theme is due to LESPERANCE AND KALBFLEISCH (1992). It is called the *intra simplex direction method*. The set of all local minima $\{\theta_1, \dots, \theta_m\}$ of $D_\phi(f_\theta - f; f)$, where D_ϕ is negative, is determined and the optimal convex combination of the current iterate and all vertices $f_{\theta_1}, \dots, f_{\theta_m}$ is the new iterate. This final step is to minimize a convex function in the variables $\epsilon_1, \dots, \epsilon_m$ under the constraint $0 \leq \sum_{i=1}^m \epsilon_i \leq 1$.

The aforementioned algorithm proposed by SIMAR (1976) and further studied in BÖHNING (1982), sticks to the original idea of picking one θ corresponding to a profitable direction. The second step differs from those indicated above. Denote by S_f the set of support points of the mixing measure corresponding to a function $f \in C$. Then, given $\hat{\theta}$, the next iterate is given by

$$g = \operatorname{argmin}_{h \in C(f)} \phi(h), \quad \text{where} \quad C(f) = \{h \in C : S_h \subset S_f \cup \{\hat{\theta}\}\}.$$

It is to be noted that support points can (and usually do) vanish during this second step. Under certain conditions, BÖHNING (1982) proves convergence of this algorithm.

In section 3 we revisit Simar’s algorithm and propose an extension of it that can deal with the case where C is the *convex cone* rather than *convex hull* generated by \mathcal{F} . This is convenient for the examples we consider. Moreover, we will introduce an algorithm that is closely related to Simar’s algorithm: the *support reduction algorithm*.

3 Support reduction and Simar's algorithm

In Simar's original algorithm, two optimization problems have to be solved. The first is to minimize the (usually nonconvex) function $D_\phi(f_\theta - f; f)$ in θ . The second is to minimize ϕ over the convex set of functions that is generated by finitely many functions from \mathcal{F} . In many examples (including the examples considered here), this second step gets more tractable if we were allowed to minimize over the *convex cone* generated by these finitely many functions in \mathcal{F} . In this section we therefore consider our function class \mathcal{F} and the convex cone C generated by it:

$$C = \text{cone}(\mathcal{F}) = \left\{ g = \int_{\Theta} f_\theta d\mu(\theta) : \mu \text{ positive finite measure on } \Theta \right\}.$$

As will be seen in section 6 and 7, our two examples fit within this framework of minimizing ϕ over the convex cone generated by a set of functions. Assumption **A2** is now replaced by the following.

Assumption A2': the function ϕ has the property that for each $f \in C = \text{cone}(\mathcal{F})$ and $g = \int_{\Theta} f_\theta d\mu_g(\theta) \in C$,

$$D_\phi(g; f) = \int D_\phi(f_\theta; f) d\mu_g(\theta). \quad (3.7)$$

Remark. Suppose that h_1 and h_2 are such that for a small positive ϵ , $f + \epsilon h_i \in C$ for $i = 1, 2$. Then, since C is convex, we have that $f + \frac{1}{2}\epsilon(h_1 + h_2) \in C$, and $D_\phi(\cdot; f)$ is well defined at h_1 , h_2 and $h_1 + h_2$. Assumption **A2'** then implies the following linearity property:

$$D_\phi(h_1 + h_2; f) = \int D_\phi(f_\theta; f) d\mu_{h_1 + h_2}(\theta) = \int D_\phi(f_\theta; f) d(\mu_{h_1} + \mu_{h_2})(\theta) = D_\phi(h_1; f) + D_\phi(h_2; f). \quad (3.8)$$

Remark. Assumption **A2'** implies **A2** for $g \in \text{conv}(\mathcal{F})$. Indeed, take $g = \int_{\Theta} f_\theta d\mu_g(\theta) \in \text{conv}(\mathcal{F})$, meaning that μ_g is a probability measure. Then we have, also using (3.8),

$$\begin{aligned} D_\phi(g - f; f) &= D_\phi(g; f) - D_\phi(f; f) = \int D_\phi(f_\theta; f) d\mu_g(\theta) - D_\phi(f; f) \\ &= \int D_\phi(f_\theta; f) - D_\phi(f; f) d\mu_g(\theta) = \int D_\phi(f_\theta - f; f) d\mu_g(\theta). \end{aligned}$$

Let us formulate a result for a generated cone analogous to lemma 2.2.

Lemma 3.1 *Let $C = \text{cone}(\mathcal{F})$ and ϕ satisfy **A1** and **A2'**. Suppose that the measure $\mu_{\hat{f}}$ in $\hat{f} = \int_{\Theta} f_\theta d\mu_{\hat{f}}(\theta)$ has finite support. Then*

$$\hat{f} = \text{argmin}_{f \in C} \phi(f) \quad \text{if and only if} \quad D_\phi(f_\theta; \hat{f}) \begin{cases} \geq 0 & \text{for all } \theta \in \Theta \\ = 0 & \text{for all } \theta \in \text{supp}(\mu_{\hat{f}}). \end{cases} \quad (3.9)$$

Proof: If $\hat{f} = \text{argmin}_{f \in C} \phi(f)$, then we have by **A1** that

$$D_\phi(\hat{f}; \hat{f}) = \lim_{\epsilon \rightarrow 0} \epsilon^{-1} \left(\phi((1 + \epsilon)\hat{f}) - \phi(\hat{f}) \right) = 0.$$

Hence, by (3.8) and lemma 2.1, we have for all $\theta \in \Theta$

$$D_\phi(f_\theta; \hat{f}) = D_\phi(f_\theta - \hat{f} + \hat{f}; \hat{f}) = D_\phi(f_\theta - \hat{f}; \hat{f}) + D_\phi(\hat{f}; \hat{f}) = D_\phi(f_\theta - \hat{f}; \hat{f}) \geq 0 \quad (3.10)$$

In view of property (3.7), we have

$$0 = D_\phi(\hat{f}; \hat{f}) = \int D_\phi(f_\theta; \hat{f}) d\mu_{\hat{f}}(\theta).$$

In the presence of the inequalities in (3.10) we therefore have that $D_\phi(f_\theta; \hat{f}) = 0$ on the support of $\mu_{\hat{f}}$ necessarily.

Conversely, if \hat{f} satisfies the (in)equalities given in (3.9) above, we have for any $f \in C$ that

$$\phi(f) - \phi(\hat{f}) \geq D_\phi(f - \hat{f}; \hat{f}) = D_\phi(f; \hat{f}) = \int D_\phi(f_\theta; \hat{f}) d\mu_f(\theta) \geq 0.$$

□

Remark. The assumption that the support of $\mu_{\hat{f}}$ is finite seems to be restrictive and unnatural. However, there are many problems (including our examples) where this is true. Of course, if Θ is finite it is trivially true (this e.g. covers interval censoring problems). Moreover, maximum likelihood estimators in mixture models usually have this property (LINDSAY (1995), theorem 18, section 5.2).

Below we give the pseudo code for Simar's algorithm constructed for a cone and also for the support reduction algorithm we propose. In fact, as will be seen below, the support reduction algorithm is Simar's algorithm where one substep is not completely followed till the end.

Basic Simar- and support reduction algorithm for a cone

Input:

$\eta > 0$: accuracy parameter;
 $\theta^{(0)} \in \Theta$: starting value;
 $f = \operatorname{argmin}_{g \in C: S_g = \{\theta^{(0)}\}} \phi(g);$

begin

while $\min_{\theta \in \Theta} D_\phi(f_\theta; f) < -\eta$ **do**

begin

$\hat{\theta} := \operatorname{argmin}_{\theta \in \Theta} D_\phi(f_\theta; f);$

$S^* := S_f \cup \{\hat{\theta}\};$

$f := \operatorname{argmin}_{g \in C: S_g \subset S^*} \phi(g);$ (*Simar*)

$f := \operatorname{argmin}_{g \in C: S_g \subset_s S^*} \phi(g);$ (*Support reduction*)

end;

end.

The meaning of ' \subset_s ' will become clear in the sequel. For both algorithms, there are two finite dimensional optimizations that have to be performed. The first one is over Θ . In general

the function $\theta \mapsto D_\phi(f_\theta; f)$ is nonconvex and minimizing such a function is usually difficult. Hence, in each setting one should try to take advantage of the specific features of that problem to attack this first optimization problem. Usually one can restrict the minimization to a bounded subset of Θ and use a fine (finite) grid in this subset instead of the whole set Θ . Then the minimization reduces to finding the minimal element in a (long) vector. After that, it is possible to ‘leave the grid’ in a way as described in section 5. Sometimes (e.g. when computing the ML estimator of a distribution function based on interval censored observations) it is even possible to select a finite subset of Θ , based on the data, such that the minimizer of ϕ over C is contained in the convex hull of the corresponding finitely many generators. In subsequent sections, we will address this matter more specifically in the examples.

The second optimization in the algorithm is over a convex cone that is spanned by finitely many functions f_θ in \mathcal{F} . Lemma 3.1 gives a characterization of such a function (applying the lemma to the finite subset S^* of Θ instead of Θ itself). We propose the following general way of solving this finite dimensional constrained optimization problem in Simar’s algorithm. In passing it will become clear what the support reduction algorithm does.

Given the current iterate and the new support point $\hat{\theta}$, consider the *linear space* L spanned by the finitely many functions $\{f_\theta : \theta \in S^*\}$:

$$L_{S^*} = \left\{ g = \int_{\Theta} f_\theta d\sigma(\theta) : \sigma \text{ is a finite signed measure on } S^* \right\},$$

and determine

$$f_u^{(0)} = \operatorname{argmin}_{g \in L_{S^*}} \phi(g) = \int f_\theta d\sigma_{f_u, S^*}(\theta) = \sum_{\theta \in S^*} f_\theta \sigma_{f_u, S^*}(\{\theta\}).$$

We assume ϕ has a smooth convex extension to the space L_{S^*} . In our examples and many others this is certainly the case. This optimization corresponds to finding a solution of a finite system of equations. Of course, $f_u^{(0)}$ will in general not be an element of C , since certain coefficients $\sigma_{f_u^{(0)}}(\{\theta\})$ may be negative. Nevertheless we can *always* move from f towards $f_u^{(0)}$ and stay within the class C initially. This is a consequence of the fact that the coefficient $\sigma_{f_u^{(0)}}(\{\hat{\theta}\})$ of $f_{\hat{\theta}}$ in $f_u^{(0)}$ will be *strictly* positive. Indeed,

$$0 > \lim_{\epsilon \downarrow 0} \epsilon^{-1} (\phi(f + \epsilon f_u^{(0)}) - \phi(f)) = \int D_\phi(f_\theta; f) d\sigma_{f_u^{(0)}}(\theta) = \sigma_{f_u^{(0)}}(\{\hat{\theta}\}) D_\phi(f_{\hat{\theta}}; f)$$

and $D_\phi(f_{\hat{\theta}}; f) < 0$ by choice of $\hat{\theta}$. If $f_u^{(0)} \in C$ then take this as next iterate. Otherwise define

$$\hat{\lambda} = \max\{l \in (0, 1] : f + l(f_u^{(0)} - f) \in C\} = \min_{\theta \in S^* : \sigma_{f_u, S^*}(\{\theta\}) < 0} (1 - \sigma_{f_u^{(0)}}(\{\theta\}) / \sigma_f(\{\theta\}))^{-1} \quad (3.11)$$

and take as next iterate the function $f + \hat{l}(f_u^{(0)} - f)$ and delete the support point $\check{\theta} \in S^*$ where the minimum in the expression on the right hand side of (3.11) is attained from the support set:

$$S^{*(1)} = S^* \setminus \{\check{\theta}\}.$$

Then compute the next unrestricted minimizer

$$f_u^{(1)} = \operatorname{argmin}_{g \in L_{S^{*(1)}}} \phi(g).$$

If this function differs from the current iterate, again a step of positive length can be made in this direction, since for all $\theta \in S^{*(1)}$, $\sigma_{f_u^{(1)}}(\{\theta\}) > 0$. If we can go all the way to $f_{u,S^*}^{(1)}$, stop the iteration, and else delete the support point as it was done in the first step. This deletion of support points can be continued until we get a subset $S^{*(j)} \subset S^*$ and a function $f_{u,S^*}^{(j)} \in C$ with support set $S^{*(j)}$ such that

$$D_\phi(f_\theta; f_{u,S^*}^{(j)}) = 0 \text{ for all } \theta \in S^{*(j)}.$$

The specific set $S^{*(j)}$ obtained in this way as subset of S^* is denoted by ' \subset_s ', and this gives the next iterate in the support reduction algorithm. Note that the function ϕ is decreased all the way during the iterations of this substep.

For Simar's algorithm, one should check for the points in $S^* \setminus S^{*(j)}$ whether the value ϕ can be improved upon by adding such points to the current support. The natural thing to do then is to take the value of $\theta \in S^{*(j)}$ where $D_\phi(f_\theta; f)$ is minimal and add this to the support. In the support reduction algorithm we skip the adding of deleted points from S^* and allow the next support point to be chosen without restriction from the whole set Θ .

Let us summarize the steps sketched above to determine $f := \operatorname{argmin}_{g \in C: S_g \subset_s S^*} \phi(g)$ in pseudo code.

Support reduction step

Input:

$f^{(0)} = f \in C$: minimizer of ϕ over subset of C consisting of functions with same support S_f ;
 $S^{*(0)} = S^* = S_f \cup \{\hat{\theta}\}$: finite set of support points;
 $j := 0$;

begin

$$f_u^{(j)} = \operatorname{argmin}_{g \in L_{S^{*(j)}}} \phi(g);$$

while $f_u^{(j)} \notin C$ **do**

begin

$$j := j + 1;$$

$$\check{\Theta} = \{\theta \in S^* : \sigma_{f_u^{(j-1)}}(\{\theta\}) < 0 \text{ and } \sigma_{f_u^{(j-1)}}(\{\theta\})/\sigma_{f^{(j-1)}}(\{\theta\}) \text{ is minimal}\};$$

$$\hat{\lambda} = (1 - \sigma_{f_u^{(j-1)}}(\{\theta\})/\sigma_{f^{(j-1)}}(\{\theta\}))^{-1} \text{ for some } \theta \in \check{\Theta};$$

$$f^{(j)} = f^{(j-1)} + \hat{\lambda}(f_u^{(j-1)} - f^{(j-1)});$$

$$S^{*(j)} := S^{*(j-1)} \setminus \check{\Theta};$$

$$f_u^{(j)} = \operatorname{argmin}_{g \in L_{S^{*(j)}}} \phi(g);$$

end;

$f := f_u^{(j)} \in C$: minimizer of ϕ over subset of C consisting of
functions with same support $S_f = S^{*(j)} \subset_s S^*$;

end.

We now see that the basic building stone of the algorithm is an *unrestricted minimization* of the function ϕ . As we will see in the sections 6 and 7, there are efficient algorithms to solve this kind of optimization problems in specific situations.

Before applying the algorithm to concrete problems, let us consider the convergence issue. The theorem below (the proof of which is inspired by BÖHNING (1982)) states that the algorithms considered in this section indeed converge to the solution of the optimization problem. To get this, we need one additional condition on the function ϕ . This condition is needed to guarantee that a strictly negative value of $D_\phi(f_\theta; f)$ for some θ means that the next iterate will have some minimal decrease in ϕ -value.

Assumption A3: For any specific starting function $f^{(0)} \in C$ with $\phi(f^{(0)}) < \infty$, there exists an $\bar{\epsilon} \in (0, 1]$ such that for all $f \in C$ with $\phi(f) < \phi(f^{(0)})$ and $\theta \in \Theta$, the following implication holds:

$$D_\phi(f_\theta - f; f) \leq -\delta < 0 \Rightarrow \phi(f + \epsilon(f_\theta - f)) - \phi(f) \leq -\frac{1}{2}\epsilon\delta \quad \text{for all } \epsilon \in (0, \bar{\epsilon}]$$

We will see that this assumption holds for the problems we will address in subsequent sections.

Theorem 3.1 *Denote by f_n a sequence generated by one of the algorithms introduced here. Then, under the assumptions A1, A2' and A3 we have that $\phi(f_n) \downarrow \phi(\hat{f})$ as $n \rightarrow \infty$.*

Proof: Since we have for each n that

$$f_n = \operatorname{argmin}_{f \in C: S_f = S_{f_n}} \phi(f),$$

we have by assumption A1 that $D_\phi(f_n; f_n) = 0$. Hence, by (3.8), we have for all $n \geq 0$

$$D_\phi(f_\theta - f_n; f_n) = D_\phi(f_\theta; f_n) \quad \text{for all } \theta \in \Theta.$$

Since $\phi(f_n)$ is a bounded and decreasing sequence of real numbers, it decreases to a limit. Assume for the moment that $\phi(f_n) \downarrow \phi^* = \phi(\hat{f}) + \delta > \phi(\hat{f})$ for some $\delta > 0$. We will extract a contradiction.

Take θ_n such that $D_\phi(f_{\theta_n}; f_n) \leq \frac{1}{2} \inf_{\theta \in \Theta} D_\phi(f_\theta; f_n)$. Then we get

$$\begin{aligned} D_\phi(f_{\theta_n} - f_n; f_n) &= D_\phi(f_{\theta_n}; f_n) \leq \frac{1}{2} \inf_{\theta \in \Theta} D_\phi(f_\theta; f_n) \leq \int \frac{1}{2} D_\phi(f_\theta; f_n) d\mu_{\hat{f}}(\theta) \\ &= \frac{1}{2} D_\phi(\hat{f} - f_n; f_n) \leq \frac{1}{2} (\phi(\hat{f}) - \phi(f_n)) \leq \frac{1}{2} (\phi(\hat{f}) - \phi^*) = -\delta/2 \end{aligned} \quad (2.12)$$

Again by monotonicity of $\phi(f_n)$, we have that $\phi(f_n) \leq \phi(f^{(0)})$ for all n , and assumption A3 gives

$$\phi(f_{n+1}) \leq \phi(f_n + \bar{\epsilon}(f_{\theta_n} - f_n)) \leq \phi(f_n) - \frac{1}{4}\bar{\epsilon}\delta \quad \text{for all } n. \quad (3.13)$$

This contradicts the fact that $\phi(f_n)$ converges. \square

In view of the convergence proof, there are some adaptations of the algorithms that do not destroy the convergence property of the algorithm. The first adaptation has to do with the choice of the most promising vertex. If the function D_ϕ on Θ is replaced by a function

$$\tilde{D}_\phi(f_\theta; f) = w(\theta)D_\phi(f_\theta; f)$$

where w is some strictly positive weight function on Θ such that

$$0 < \underline{w} \leq w(\theta) \leq \overline{w} < \infty \quad \text{for all } \theta \in \Theta.$$

Equation (3.12) would then change to

$$D_\phi(f_{\theta_n} - f_n; f_n) = D_\phi(f_{\theta_n}; f_n) \leq (\underline{w}/\overline{w}) \inf_{\theta \in \Theta} D_\phi(f_\theta; f_n) \leq -(\underline{w}/\overline{w})\delta,$$

and the argument goes through with δ replaced by $(\underline{w}/\overline{w})\delta$. Similarly, **A3** will also hold for \tilde{D}_ϕ if it holds for D_ϕ . In section 4 we will use this idea to define an alternative directional derivative function.

The second adaptation is the following. If it is possible after *reduction by deletion* of support points to do an extra step of *reduction by replacing two support points by a third* or move a support point slightly without increasing the function ϕ , this will not prevent the algorithm from converging. This immediately follows from (3.13). Indeed, if we replace the iterate f_n that would be obtained by the original method by \tilde{f}_n which satisfies

$$\tilde{f}_n = \operatorname{argmin}_{f \in C: S_f = S_{\tilde{f}_n}} \phi(f) \text{ and } \phi(\tilde{f}_n) \leq \phi(f_n),$$

the inequality (3.13) also holds for \tilde{f}_{n+1} instead of f_{n+1} and the proof goes through. This adaptation of the algorithm will be discussed more elaborately in section 5.

4 Alternative directional derivative

Consider a quadratic objective function ϕ_q on $C = \operatorname{cone}(\{f_\theta : \theta \in \Theta\})$. The objective function in the LS estimation context is quadratic automatically and in section 7 we will use a Newton algorithm to solve the ML estimation problem. In that algorithm a quadratic approximation of the objective function is minimized during each iteration.

The function ϕ_q is quadratic in f . Hence, along line segments in the linear space spanned by the functions $f_{\theta_1}, \dots, f_{\theta_p}$, the function is also quadratic as a function of one variable. Along such segments (or lines), the function ϕ_q can therefore be minimized explicitly. Given a ‘current iterate’ g in the algorithm, we consider for each $\theta \in \Theta$ the following function (alternative choice is to take $f_\theta - g$ instead of f_θ):

$$\epsilon \mapsto \phi_q(g + \epsilon f_\theta) - \phi_q(g) = c_1(\theta, g)\epsilon + \frac{1}{2}\epsilon^2 c_2(\theta).$$

Typically, $c_2 > 0$, so that

$$\hat{\epsilon} = \hat{\epsilon}_\theta = \operatorname{argmin}_\epsilon \phi_q(g + \epsilon f_\theta) = -\frac{c_1(\theta, g)}{c_2(\theta)}$$

is the optimal move along the line connecting g and $g + f_\theta$.

In order to have descent direction, we only consider points θ where $c_1(\theta, g) < 0$. In that case, $\hat{\epsilon} > 0$. As new vertex, we then define

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta: c_1(\theta) < 0} \phi_q(g + \hat{\epsilon}(\theta)f_\theta) = \operatorname{argmin}_{\theta \in \Theta: c_1(\theta) < 0} - \frac{c_1(\theta, g)^2}{2c_2(\theta)} = \operatorname{argmin}_{\theta \in S} \frac{c_1(\theta, g)}{\sqrt{c_2(\theta)}}.$$

5 A ‘gridless’ implementation

For a practical implementation of the step of selecting a new support point, we propose to fix a fine grid Θ_δ in Θ and run the whole algorithm with Θ_δ instead of Θ . Having a precise approximation of the minimizer f_{grid} of ϕ over this finite dimensional cone, one can make the algorithm ‘gridless’ by fine tuning in the support points. This can be done by augmenting a step at each iteration in the spirit of the second remark after theorem 3.1.

Write f for the current iterate (at the first ‘fine tuning step’, this is f_{grid}) and define

$$\tau(h_1, h_2, \dots, h_p) = \tau(h_1, h_2, \dots, h_p; f) = \phi\left(\sum_{i=1}^p \alpha_i f_{\theta_i + h_i}\right) - \phi\left(\sum_{i=1}^p \alpha_i f_{\theta_i}\right)$$

with $\alpha_1, \dots, \alpha_p$ fixed and $h = (h_1, \dots, h_p)^T$ varying over a neighborhood of zero in \mathbb{R}^p . The function τ represents the value of the objective function if the masses α_i are fixed and the current support points are shifted a bit. Abusing notation slightly, write

$$f_h = \sum_{i=1}^p \alpha_i f_{\theta_i + h_i}$$

and note that for the least squares objective function (under mild smoothness assumptions on the parameterization of f_θ)

$$\frac{\partial \tau}{\partial h_i}(h_1, h_2, \dots, h_p) = \alpha_i \int \dot{f}_{\theta_i + h_i}(x) f_h(x) dx - \alpha_i \int \dot{f}_{\theta_i + h_i}(x) d\mathbb{F}_n(x) \quad (5.14)$$

and for the maximum likelihood objective function

$$\frac{\partial \tau}{\partial h_i}(h_1, h_2, \dots, h_p) = -\alpha_i \int \frac{\dot{f}_{\theta_i + h_i}(x)}{f_h(x)} d\mathbb{F}_n(x). \quad (5.15)$$

In particular, note that

$$\frac{\partial \tau}{\partial h_i}(0) = \alpha_i \frac{d}{d\theta} D_\phi(f_\theta; f)|_{\theta=\theta_i}$$

for both objective functions. Hence, the partial derivatives of τ at zero are visualized in the graph of $\theta \mapsto D_\phi(f_\theta; f)$ for both objective functions. Qualitatively, the interpretation of the partial derivatives of τ is that if $\frac{\partial}{\partial h_i} \tau(0) < 0$, shifting the support point θ_i slightly to the right will result in a decrease of the objective function. For the moment, fix $h \in \mathbb{R}^p$ with $\|h\|_2 = 1$ and consider the function

$$\mu_h(\epsilon) = \tau(\epsilon h)$$

on an interval $[0, \epsilon_0]$ for some small $\epsilon_0 > 0$. Note that $\tau_h(0) = 0$. Then (again under mild smoothness assumptions) the derivative of μ_h is given by

$$\mu'_h(\epsilon) = h^T \nabla \tau(\epsilon h)$$

where $\nabla \tau(\epsilon h)$ is given either by (5.14) or (5.15), depending on the objective function. Taking $\epsilon = 0$, we see that the ‘most promising’ direction to move, is the direction $-\nabla \tau(0)$, the direction of steepest descent. From now on take this direction. The aim is now to move the support points in this direction to get a sufficient decrease in the objective function. This means that μ_h is to be minimized as a function of $\epsilon \in (0, \epsilon_0)$, or at least a value ϵ has to be determined such that $\mu_h(\epsilon)$ is negative. Note that the function μ_h is *nonconvex* in general. We determine the step length by the method of regula falsi on the derivative μ_h . At zero this function is zero. Define $\epsilon_l = 0$ and $\epsilon_u = \epsilon_0$. If $\mu_h(\epsilon_u) < 0$ then take this $\epsilon = \epsilon_0$. Otherwise proceed as follows.

$$\epsilon_n = \frac{\epsilon_l \mu'_h(\epsilon_u) - \epsilon_u \mu'_h(\epsilon_l)}{\mu'_h(\epsilon_u) - \mu'_h(\epsilon_l)}.$$

If $\mu'_h(\epsilon_n) > 0$ define $\epsilon_u = \epsilon_n$ whereas if $\mu'_h(\epsilon_n) < 0$ define $\epsilon_l = \epsilon_n$. This process can be iterated till $\mu'_h(\epsilon_n)$ is sufficiently small in absolute value. This regula falsi method comes up with a stationary point of μ_h . If the $\mu_h(\epsilon_n)$ is positive, the line search procedure should be repeated with $\epsilon_0 = c\epsilon_n$ for some $0 < c < 1$ (usually close to one). In our experience this step is hardly ever necessary, but conceptually it is needed. The procedure will (in case $\epsilon \neq \epsilon_0$) lead to a stationary point of μ_h with $\mu_h(\epsilon) < \mu_h(0) = 0$. Actually, ϵ will usually correspond to the smallest local minimum of μ_h .

Next, define

$$\bar{f} := \sum_{i=1}^p \alpha_i f_{\theta_i + \epsilon h_i}.$$

The new iterate f is finally obtained by minimizing ϕ over the cone generated by $\{f_{\theta_i + \epsilon h_i} : 1 \leq i \leq p\}$. This function satisfies the conditions needed at the beginning of the just described ‘fine tuning’ step. Hence, it can be iterated till the norm of $\mu'_h(0)$ is sufficiently small.

6 LS estimation of a convex density

In this section we study the problem of computing the least squares estimate of a convex and decreasing density on $[0, \infty)$. In GROENEBOOM, JONGBLOED AND WELLNER (2001B), it is shown that the (uniquely defined) minimizer of the convex function ϕ over $\text{conv}(\mathcal{F})$ is the same as the minimizer of ϕ over $\text{cone}(\mathcal{F})$. It is also shown that there only functions f_θ with $\theta \in [x_1, K]$ for some $K < \infty$ have to be considered in the optimization, since the optimal function has no change of slope at a location to the left of x_1 and has compact support. Hence we are in the situation of section 3. Moreover, we have

$$D_\phi(f_\theta; f) = \int_0^\infty f_\theta(x) f(x) dx - \int_0^\infty f_\theta(x) d\mathbb{F}_n(x) = \frac{2}{\theta^2} (H(\theta; f) - Y_n(\theta))$$

where

$$H(\theta; f) = \int_{x=0}^{\theta} \int_0^x f(y) dy dx \text{ and } Y_n(\theta) = \int_0^{\theta} \mathbb{F}_n(x) dx;$$

here we use the same notation as in GROENEBOOM, JONGBLOED AND WELLNER (2001B). Note that the assumptions **A1**, **A2'** and **A3** are satisfied in this situation. For **A3** note that

$$\phi(f + \epsilon(f_{\theta} - f)) = \phi(f) + \epsilon D_{\phi}(f_{\theta} - f; f) + \frac{1}{2}\epsilon^2 \int_0^{\infty} (f_{\theta}(x) - f(x))^2 dx \quad (6.16)$$

and that for $\theta \in [x_1, K]$

$$\int_0^{\infty} (f_{\theta}(x) - f(x))^2 dx = \frac{4}{3\theta} - \frac{4}{\theta^2} H(\theta; f) + \int_0^{\infty} f(x)^2 dx \leq M$$

for some big finite M not depending on θ .

Let us now consider the support reduction algorithm. To start this algorithm, we fix a starting value $\theta^{(0)} > x_n$. Then we determine the function cf_{θ_0} minimizing ϕ as function of $c > 0$. To this end we need the value c that minimizes

$$c \mapsto \phi(cf_{\theta^{(0)}}) = \frac{1}{2}c^2 \int_0^{\infty} f_{\theta^{(0)}}(x)^2 dx - c \int_0^{\infty} f_{\theta^{(0)}}(x) d\mathbb{F}_n(x) = \frac{2c^2}{3\theta^{(0)}} - \frac{2c(\theta^{(0)} - \bar{x}_n)}{(\theta^{(0)})^2}$$

giving $c = \frac{3}{2}(1 - \bar{x}_n/\theta^{(0)})$. If $x_n < 3\bar{x}_n$, one could also choose to take $\theta^{(0)} = 3\bar{x}_n$, so that the starting function $f^{(0)}$ would be a density.

The two main steps are minimizing $D_{\phi}(f_{\theta}; f)$ as a function of θ and minimizing the function ϕ over the space of piecewise linear functions with bend points in a finite set S^* . For the first step, we follow the line of thought given in section 4. In this example we have for all $\epsilon > 0$ that

$$\phi(f + \epsilon f_{\theta}) = \phi(f) + \epsilon c_1(\theta, f) + \frac{1}{2}\epsilon^2 c_2(\theta) \text{ with } c_1(\theta, f) = D_{\phi}(f_{\theta}; f) \text{ and } c_2(\theta) = \frac{4}{3\theta}.$$

Hence, we have as ‘alternative directional derivative’ function

$$\tilde{D}_{\phi}(f_{\theta}; f) = \frac{c_1(\theta, f)}{\sqrt{c_2(\theta)}} \simeq \sqrt{\theta} D_{\phi}(f_{\theta}; f)$$

where \simeq denotes ‘equality apart from a positive multiplicative constant’. Note that, since $w(\theta) = \sqrt{\theta}$ is strictly positive and uniformly bounded away from zero and infinity on $\Theta = [x_1, K]$, we are in the situation described below theorem 3.1. Note that $\theta \mapsto \tilde{D}_{\phi}(f_{\theta}; f)$ is continuous,

$$\tilde{D}_{\phi}(f_{\theta}; f) = 0 \text{ at } \theta = 0 \text{ and } \lim_{\theta \rightarrow \infty} \tilde{D}_{\phi}(f_{\theta}; f) = 0.$$

Hence, if $\tilde{D}_{\phi}(f_{\theta}; f) < 0$ for some θ , it attains its minimal value.

The second step in the algorithm boils down to the following procedure. Write $S^* = S_f \cup \{\hat{\theta}\} = \{\theta_1, \theta_2, \dots, \theta_m\}$ with $\theta_1 < \dots < \theta_m$ and construct a cubic spline P with knots $\{\theta_1, \theta_2, \dots, \theta_m\}$ such that

$$P(\theta) = Y_n(\theta) \text{ for all } \theta \in S^*, \quad P(0) = P'(0) = P''(\theta_m) = 0. \quad (6.17)$$

Note that the second derivative of this cubic spline minimizes the function ϕ within the class of linear splines l with knots concentrated on the set $\{\theta_1, \theta_2, \dots, \theta_m\}$ subject to the boundary constraint that $l(\theta_m) = 0$. This follows by setting the derivatives of ϕ in the directions f_{θ_j} equal to zero, i.e. $D_\phi(f_{\theta_j}; f) = 0$.

Figure 1 shows the results of the SR algorithm based on a sample of size 500 from the standard exponential distribution. The solution on an equidistant grid in $[0, 3x_{(n)}] = [0, 16.5]$ consisting of 1000 points was obtained after 33 iterations. Furthermore, we used accuracy parameter $\eta = 10^{-10}$.

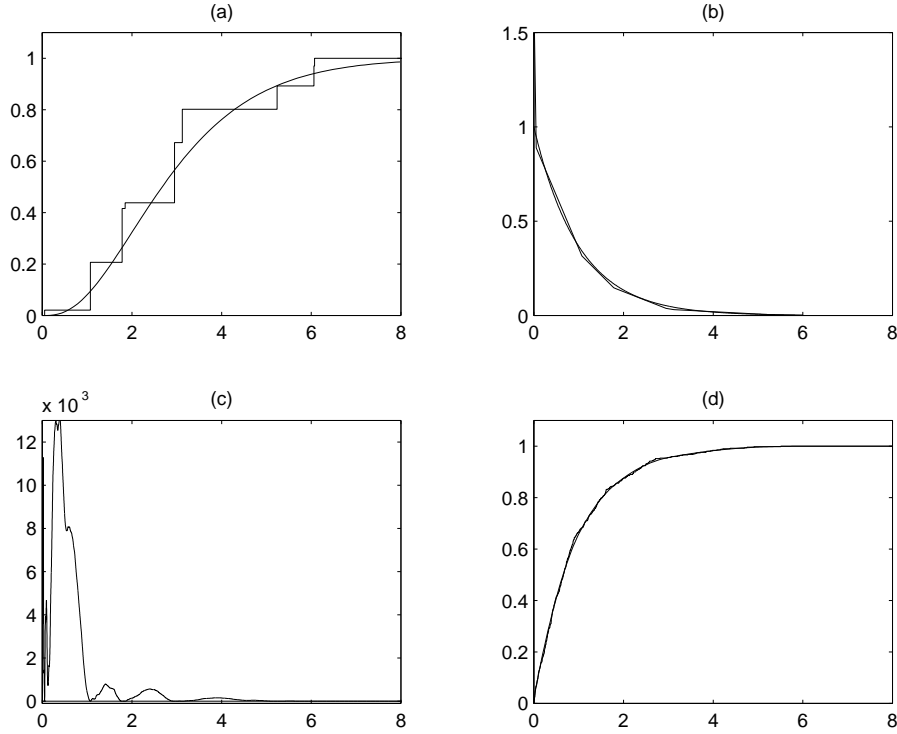


Figure 1: (a) LS estimate of the mixing distribution with the true Gamma (3) mixing distribution; (b) LS estimate of the mixture density with the true density; (c) the (alternative) directional derivative function evaluated at the LS estimate and (d) LS estimate of the mixture distribution with the empirical distribution function of the data. All pictures are based on a sample of size $n = 500$ from the standard exponential distribution.

7 ML estimation in Gaussian deconvolution

In order to apply the support reduction algorithm of section 3, the setting of Example 2 is not appropriate since the minimization there has to be performed over the convex hull of the functions f_θ instead of the convex cone generated by them. Contrary to the situation of section 6, the minimizer of ϕ over the cone does not exist (given a function f with $\phi(f) < 0$, the function ϕ applied to $c \cdot f$ for $c > 0$ tends to minus infinity). To get a well posed

optimization problem over the convex cone so that its solution is the minimizer of ϕ over the convex hull of \mathcal{F} , we have to relax the constraint that the solution has to be a probability density. The new objective function then becomes

$$\phi(f) = - \int \log f(x) d\mathbb{F}_n(x) + \int f(x) dx.$$

In principle, the support reduction algorithm can be applied directly to the thus obtained optimization problem. However, we observed that a Newton-type procedure (based on the support reduction algorithm) worked significantly better than the direct application of the support reduction algorithm. We describe this Newton procedure here. Write \bar{f} for the *current iterate*.

Note that

$$\phi(f) - \phi(\bar{f}) = - \int \log \left(1 + \frac{f(x) - \bar{f}(x)}{\bar{f}(x)} \right) d\mathbb{F}_n(x) + \int f(x) - \bar{f}(x) dx$$

For $(f - \bar{f})/\bar{f}$ small, we get the following quadratic approximation of ϕ at \bar{f} , using the second order Taylor approximation of the logarithm at 1

$$\int \frac{1}{2} \left(\frac{f(x) - \bar{f}(x)}{\bar{f}(x)} \right)^2 - \frac{f(x) - \bar{f}(x)}{\bar{f}(x)} d\mathbb{F}_n(x) + \int f(x) - \bar{f}(x) dx$$

Ignoring terms that do not depend on f , we define the following local objective function

$$\phi_q(f) = \phi_q(f; \bar{f}) = \int f(x) dx + \int \frac{1}{2} \left(\frac{f(x)}{\bar{f}(x)} \right)^2 - 2 \frac{f(x)}{\bar{f}(x)} d\mathbb{F}_n(x)$$

This quadratic function can be minimized over the (finitely generated) cone using the support reduction algorithm, yielding

$$\bar{f}_q = \operatorname{argmin}\{\phi_q(f; \bar{f}) : f \in \operatorname{cone}(f_\theta : \theta \in \Theta_\delta)\}$$

The next iterate is then obtained as $\bar{f} + \lambda(\bar{f}_q - \bar{f})$ (λ chosen appropriately to assure monotonicity of the algorithm).

This method is used to solve the (finite dimensional) optimization problem over the cone of functions generated by $\{f_\theta : \theta \in \Theta_\delta\}$. After this, the fine tuning in support points (leaving the prespecified grid) is performed as described in section 5.

During the Newton iterations to obtain the solution to the finite dimensional problem as well as in the fine tuning step following it, quadratic optimization problems of the type find

$$\operatorname{argmin}\{\phi_q(f) : f \in \operatorname{cone}(f_\theta : \theta \in S)\}$$

are solved for some *finite* set S . Starting from an initial value, say g (the natural candidate for this will be obvious from the context; usually it has only a few active vertices), the support reduction algorithm consists of two steps that are iterated:

- 1) Find new support point

- 2) Do finite dimensional constrained optimization using iterative unconstrained minimizations.

Step 1. In the notation of section 4, we have

$$c_1(\theta, g) = \int f_\theta(x) dx - 2 \int \frac{f_\theta}{\bar{f}}(x) d\mathbb{F}_n(x) + \int \frac{g f_\theta}{\bar{f}^2}(x) d\mathbb{F}_n(x) \text{ and } c_2(\theta) = \int \frac{f_\theta^2}{\bar{f}^2}(x) d\mathbb{F}_n(x).$$

Hence, the new vertex is given by

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta_\delta} \frac{c_1(\theta, g)}{\sqrt{c_2(\theta)}}$$

Step 2. During this step, given a support set $\{\theta_1, \dots, \theta_p\}$, we should find a subset S of $\{\theta_1, \dots, \theta_p\}$ with associated optimal f such that f minimizes ϕ_q over the linear space generated by the functions $\{f_\theta : \theta \in S\}$ and, moreover, has only scalars $\alpha_j > 0$ in the representation

$$f = \sum_{\theta_j \in S} \alpha_j f_{\theta_j}$$

The basic step in finding S and f is to minimize, without restrictions on α_j , the quadratic function

$$\begin{aligned} \psi(\alpha_1, \dots, \alpha_p) &= \phi_q\left(\sum_{\theta_j \in S} \alpha_j f_{\theta_j}\right) \\ &= \sum_{i=1}^p \alpha_i \left(\int f_{\theta_i}(x) dx - 2 \int \frac{f_{\theta_i}(x)}{\bar{f}(x)} d\mathbb{F}_n(x) \right) + \frac{1}{2} \sum_{i=1}^p \sum_{j=1}^p \alpha_i \alpha_j \int \frac{f_{\theta_i}(x) f_{\theta_j}(x)}{\bar{f}(x)^2} d\mathbb{F}_n(x) \\ &= \alpha^T \nu + \frac{1}{2} \alpha^T V \alpha \end{aligned}$$

Define the $n \times p$ -matrix Y by $Y_{ij} = f_{\theta_j}(x_i)$. Note that this matrix only depends on the values of the current vertices at the observed sample. Also define the n -vector d by $d_i = (\bar{f}(x_i))^{-1}$ and the $n \times n$ diagonal matrix D $D_{ii} = d_i$. Then $nV = Y^T D^T D Y$ and $n\nu = n_p - 2Y^T d$ (using that the vertices are in fact *probability densities*, denoting by n_p the p -vector with all elements equal to n) and the optimal $\alpha \in \mathbb{R}^p$ minimizing ψ is the solution to the following linear system of equations

$$(DY)^T DY \alpha = 2Y^T d - n_p$$

If the matrix DY has full rank p , this system has a unique solution.

Figure 2 shows the results of the SR algorithm based on a simulated dataset of size $n = 500$ where the mixing distribution is standard exponential. First it took 25 iterations to obtain the solution on an equidistant grid of size 500 in $[x_1, x_{(n)}] = [-2.47, 7.96]$. This grid-solution had eight support points. After that, 1085 steps of the fine tuning step of section 5 were taken, resulting in an estimate of the mixing distribution with five support points.

Acknowledgement: We thank Jon Wakefield for drawing our attention to Mallet's paper.

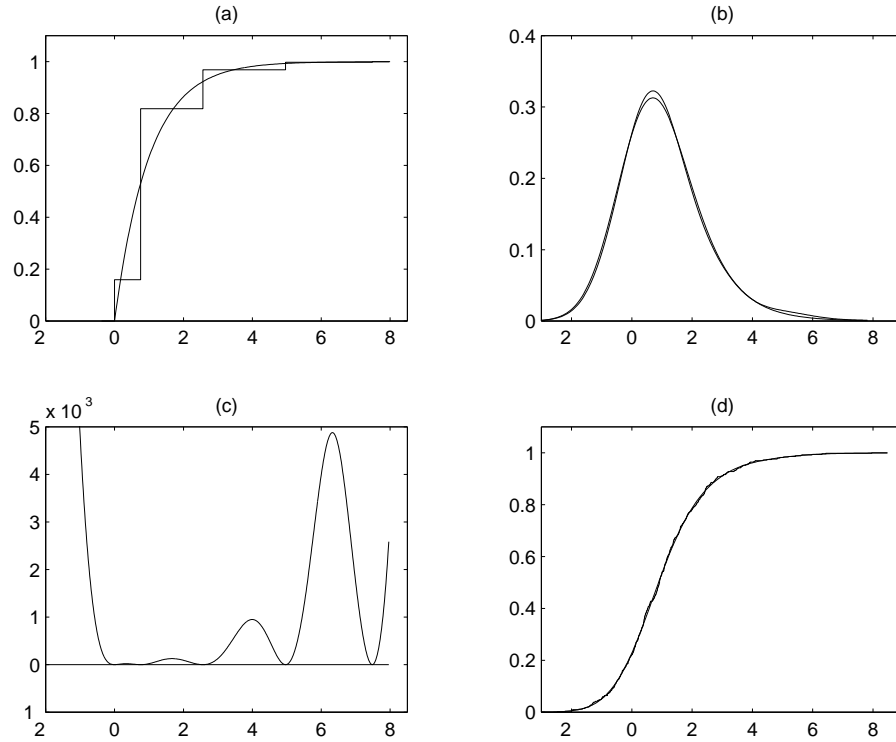


Figure 2: (a) ML estimate of the mixing distribution with the true mixing distribution; (b) ML estimate of the mixture density with the true density; (c) the (alternative) directional derivative function evaluated at the ML estimate and (d) ML estimate of the mixture distribution with the empirical distribution function of the data. All pictures are based on a sample of size $n = 500$ from the standard exponential mixture of standard normals.

References

- BÖHNING, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.* **10**, 1006–1008.
- BÖHNING, D. (1986). A vertex-exchange method in D -optimal design theory. *Metrika* **33**, 337–347.
- DEMPSTER, A.P., LAIRD, N.M. AND RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39**, 1–38.
- FEDOROV, V.V. (1972). *Theory of optimal experiments*. Academic, New York.
- GROENEBOOM, P., JONGBLOED, G., AND WELLNER, J.A. (2001A). A canonical process for estimation of convex functions: the "envelope" of integrated Brownian motion $+t^4$. *Ann. Statist.* **29**, 1620–1652.
- GROENEBOOM, P., JONGBLOED, G., AND WELLNER, J.A. (2001B). Estimation of convex functions: characterizations and asymptotic theory. *Ann. Statist.* **29**, 1653–1698.
- GROENEBOOM, P. AND WELLNER, J.A. (1992). *Information bounds and nonparametric*

- maximum likelihood estimation*. Birkhäuser, Basel.
- HAMPEL, F.R. (1987). Design, modelling and anlysis of some biological datasets. In *Design, data and analysis, by some friends of Cuthbert Daniel*, C.L. Mallows, editor, 111- 115. Wiley, New York.
- JONGBLOED, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comp. Graph. Statist.* **7**, 310–321.
- LAVEE, D., SAFRIE, U.N., AND MEILIJSO, I. (1991). For how long do trans-Saharan migrants stop over at an oasis? *Ornis Scandinavica* **22**, 33-44.
- LESPERANCE, M.L. AND KALBFLEISCH, J.D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the Americal Statistical Association* **87** 120–126.
- LINDSAY, B.G. (1995). *Mixture models: theory, geometry and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, **Vol 5**.
- MALLET, A. (1986). A maximum likelihood estimation method for random coefficient regression models. *Biometrika* **73**, 645–656.
- ROBERTSON, T., WRIGHT, F. T., DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.
- SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4**, 1200–1209.
- WELLNER, J.A. AND ZHANG, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, 779–814.
- WU, C.F. (1978). Some algorithmic aspects of the theory of optimal design. *Ann. Statist.* **6**, 1286–1301.
- WYNN, H.P. (1970). The sequential generation of D -optimum experimental designs. *Ann. Math. Statist.* **6** 1286–1301.